attempts at chain-end moves, N-1 attempts at two-bond moves and one attempt at a randomly selected, large fragment displacement. Here, "N" equals the number of amino acid residues in the protein. Before any energy computation, a test for excluded volume violations is performed, and trial conformations that would lead to steric collisions of chain units are rejected, as are conformations that would result in nonphysical distances between two consecutive side chain units.

Interaction Scheme

The interaction scheme employed in SICHO comprises short-range interactions, hydrogen bond interactions, and long-range interactions. All types of interactions have generic (*i.e.*, sequence-independent), sequence-dependent, and target (*i.e.*, resulting from superimposed short- and long-range constraints) components. Below, the generic and sequence-dependent terms are described first, followed by a description of those terms arising from the constraint contributions.

Sequence-dependent short-range interactions

The potentials were derived from the geometric statistics of known protein structures. Pairwise-specific distances between nearest neighbors, up to the fourth neighbor, along the polypeptide chain are considered. These distances depend on amino acid composition and the local chain geometry. Six bins, covering the majority of distances, including the more distant pairs, *i.e.*, the wings of the distance distribution (which are cut off at 4.8-7.9 Å) observed in proteins, have been used for all components of the short-range interactions. For a given pair of amino acid residues, the distribution of associated distances between side chain centers of mass is extracted from a statistical analysis of a structural database of non-homologous proteins (the Holm Sander PDB select database of 1501 proteins). When compared to an average distribution (ignoring sequence information), this leads to a statistical potential. The technique is similar to that employed elsewhere.[15] As schematically illustrated in Figure 4, the resulting potential could be expressed as follows:

$$E_{short} = \sum E_{12}(r^2_{i,i+1}, A_i, A_{i+1}) + \sum E_{13}(r^2_{i,i+2}, A_i, A_{i+2}) + \sum E_{14}(r^{2*}_{i,i+3}, A_{i+1}, A_{i+2}) + \sum E'_{14}(r^{2*}_{i,i+3}, A_i, A_{i+3}) + \sum E_{15}(r^2_{i,i+4}, A_{i+2}, A_{i+3}) + \sum E'_{15}(r^2_{i,i+4}, A_j, A_{i+4}) . \quad (1)$$

The summation is performed along the chain; E_{1d} refers to energy associated with interactions between the residue of interest and its d-1st neighbor down the chain. A_i denotes the amino acid identity at position i, and $r_{i,i-k}$ is the distance between residues i and i + k. The terms for the three-bond fragments include the effects of local chain chirality via a "chiral"-distance-squared term.

$$r^{2*}_{i-1,i+2} = r^2_{i-1,i-2} \, \mathrm{sign}((\mathbf{v}_{i-1} \otimes \mathbf{v}_i) \cdot \mathbf{v}_{i+1}) . \quad (2)$$

All terms are amino acid pair-specific because the presently available structural database do not support meaningful statistics for higher order terms. Thus, there is a single energy term for one-bond and two-bond fragments, and two types of binary potentials for three-bond and four-bond fragments. These sequence dependent short-range interactions also provide information about short-range packing regularities, *e.g.*, the propensities for a particular side chain arrangement on a helical surface. For simplicity, the relative scaling of all terms is preferably taken to be equal to one. This scaling generates a reasonable level and identity of secondary structure. While other scaling factors could be used, the quality of the results drops off, for example, less than native secondary structure or too much and poor backbone geometry are derived. Since there are a large number of numerical values for these short-range potentials (six components, each having 20 x 20 x 6 pair-wise values for 6-bin histograms), the data been reported[44] and are available via anonymous ftp[17].

Generic short-range conformational biases

Next, terms that do not depend on amino acid sequence are introduced into the model force field. Thus, the energy contribution from these terms depends only on specific chain geometry (regardless of protein sequence) and its magnitude is controlled by a single adjustable energetical parameter, ε_{gen}. These terms' purpose is to enforce a protein-like distribution of short-range conformations.

The first set of these terms accounts for the characteristic stiffness of polypeptide chains, which builds on the observation that there is a characteristic orientation of protein chain that could be conveniently defined by a vector orthogonal to a triangle formed by three consecutive centers of mass of the side chains. The corresponding conformational bias could be defined as follows:

$$E_{stiff} = -0.25\ \varepsilon_{gen} \Sigma\ (\mathbf{w}_i \cdot \mathbf{w}_{i+4}) \qquad (3)$$

where $\mathbf{w}_i$ is a vector orthogonal to the plane formed by the two consecutive virtual covalent bonds $\mathbf{v}_{i-1}$ and $\mathbf{v}_i$, ε_{gen} is an arbitrarily chosen energetic parameter equal to 1 k_BT in all potentials described in this section, here scaled by a factor equal to -0.25. The length of the orthogonal vectors $\mathbf{w}_i$ is about 4 lattice units, and they are also used for detection of "hydrogen bonds." The dot product in the above equation is near its maximum value for extended, β-like states and for helices. The high value of this product is significant in a majority of typical turns and loop-type local conformations. Thus, the potential provides a bias towards these relatively rigid elements of protein secondary structure.

The second generic term provides a bias towards regular arrangements of secondary structure. In a random lattice chain, the distribution of distances between the i-th and $i + 4^{th}$ bead would be unimodal and close to a Gaussian distribution. On the other hand, the corresponding distance distribution between residues in native proteins is bimodal. The shorter distance peak corresponds to helical and turn conformations, while the more diffuse, longer distance peak corresponds to extended